US009471673B1

(12) United States Patent
Sharifi et al.

(10) **Patent No.: US 9,471,673 B1**
(45) **Date of Patent: Oct. 18, 2016**

(54) **AUDIO MATCHING USING TIME-FREQUENCY ONSETS**

(75) Inventors: **Matthew Sharifi**, Zurich (CH); **Richard Francis Lyon**, Los Altos, CA (US)

(73) Assignee: **Google Inc.**, Mountain View, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 1013 days.

(21) Appl. No.: **13/418,334**

(22) Filed: **Mar. 12, 2012**

(51) **Int. Cl.**

G10L 15/00	(2013.01)
G06F 17/30	(2006.01)
G10L 15/02	(2006.01)

(52) **U.S. Cl.**
CPC ... ***G06F 17/30743*** (2013.01); *G10H 2210/051* (2013.01); *G10H 2210/061* (2013.01); *G10H 2240/141* (2013.01); *G10L 15/02* (2013.01)

(58) **Field of Classification Search**
CPC .. G06F 17/00; G06F 17/18; G06F 17/30017; G06F 17/3002; G06F 17/30023; G06F 17/30029; G06F 17/3033; G06F 17/30743; G06F 17/30761; G06F 17/30784; G06F 17/30799; G06F 21/10; G06K 9/00; G06K 9/00744; G06K 9/00758; G06K 9/62; G06T 9/002; G06T 9/005; G10H 2210/061; G10L 19/00; G10L 19/018; G10L 25/18; G10L 25/48; G10L 25/51; G10L 25/54; G11B 27/034; G11B 27/28; H04H 20/14; H04H 60/37; H04H 60/372; H04H 60/375; H04H 60/58; H04H 60/59; H04H 60/64; H04H 2201/90; H04L 2209/60; H04N 21/26603; H04N 21/4394; Y10S 707/99933; Y10S 707/99936
USPC 382/100, 181, 190, 224; 700/94; 704/231, 243, 245, 256, 270, 273, 704/E11.001, E11.002; 705/57; 707/737, 707/741, 747, 749, 752, 758, 769, 999.003, 707/999.006, E17.014, E17.101; 713/176, 713/180; 725/18, 19; 726/26
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,393,450	B1 *	5/2002	Whikehart	708/316
6,453,252	B1	9/2002	Laroche	
6,721,488	B1	4/2004	Dimitrova et al.	
7,516,074	B2	4/2009	Bilobrov	
7,809,580	B2	10/2010	Hotho et al.	
8,488,800	B2 *	7/2013	Crockett	381/56
2002/0023020	A1	2/2002	Kenyon et al.	
2002/0037083	A1 *	3/2002	Weare et al.	381/58
2005/0104654	A1 *	5/2005	Muhammad et al.	327/552
2005/0273319	A1 *	12/2005	Dittmar et al.	704/203

(Continued)

OTHER PUBLICATIONS

Project Report: Onset Detection, Mikkel Gravgaard Nielsen Aarhus University, Feb. 4, 2010 http://cs.au.dk/~grav/onset/.*
Media Hedge, "Digital Fingerprinting," White Paper, Civolution and Gracenote, 2010, http://www.civolution.com/fileadmin/bestanden/white%20papers/Fingerprinting%20-%20by%20Civolution%20and%20Gracenote%20-%202010.pdf, Last accessed Jul. 11, 2012.
Milano, Dominic, "Content Control: Digital Watermarking and Fingerprinting," White Paper, Rhozet, a business unit of Harmonic Inc., http://www.rhozet.com/whitepapers/Fingerprinting_Watermarking.pdf, Last accessed Jul. 11, 2012.

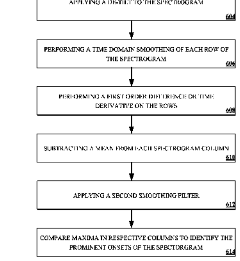
(Continued)

Primary Examiner — Paul McCord
Assistant Examiner — Alexander Eljaiek
(74) *Attorney, Agent, or Firm* — Fenwick & West LLP

(57) **ABSTRACT**

Systems and methods are provided herein relating to audio matching. Interest points that are onsets are generally very efficient in audio matching in that they are robust to multiple types of distortion. Prominent onsets can be detected within an audio signal excerpt as interest points and combined as a function of a set of interest points to form a descriptor. Descriptors associated with an audio signal excerpt that contain a set of prominent onsets as interest points can be used in matching the audio signal excerpt to an audio reference. The benefits in generating and using prominent onsets within descriptors improve the accuracy of an audio matching system.

23 Claims, 11 Drawing Sheets


600
GENERATING A SPECTROGRAM OF AN AUDIO SIGNAL EXCERPT USING SHORT TIME FOURIER TRANSFORMS (STFTs) ON OVERLAPPING WINDOWS OF THE AUDIO SIGNAL EXCERPT
602
APPLYING A DE-TILT TO THE SPECTROGRAM
604
PERFORMING A TIME DOMAIN SMOOTHING OF EACH ROW OF THE SPECTROGRAM
606
PERFORMING A FIRST ORDER DIFFERENCE OR TIME DERIVATIVE ON THE ROWS
608
SUBTRACTING A MEAN FROM EACH SPECTROGRAM COLUMN
610
APPLYING A SECOND SMOOTHING FILTER
612
COMPARE MAXIMA IN RESPECTIVE COLUMNS TO IDENTIFY THE PROMINENT ONSETS OF THE SPECTORGRAM
614

(56) **References Cited**

U.S. PATENT DOCUMENTS

2007/0282860	A1*	12/2007	Athineos et al.	707/10
2008/0300702	A1*	12/2008	Gomez et al.	700/94
2009/0012638	A1*	1/2009	Lou	700/94
2013/0139673	A1*	6/2013	Ellis et al.	84/609

OTHER PUBLICATIONS

MusicBrainz—The Open Music Encyclopedia, http://musicbrainz.org, Last accessed Apr. 12, 2012.

Shazam, http://www.shazam.com, Last accessed Apr. 19, 2012.

* cited by examiner

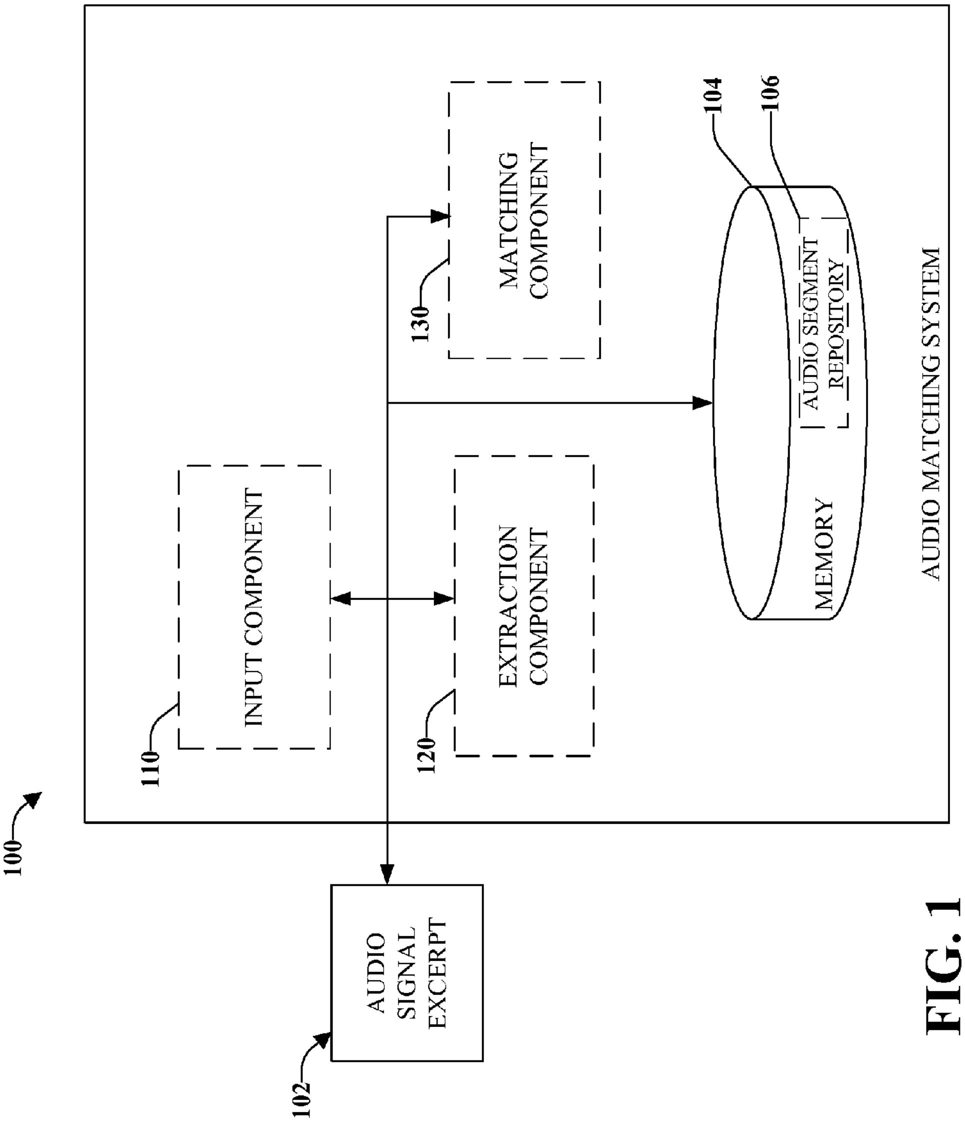

100
AUDIO MATCHING SYSTEM
110
INPUT COMPONENT
120
EXTRACTION COMPONENT
130
MATCHING COMPONENT
104
106
AUDIO SEGMENT REPOSITORY
MEMORY
102
AUDIO SIGNAL EXCERPT

FIG. 1

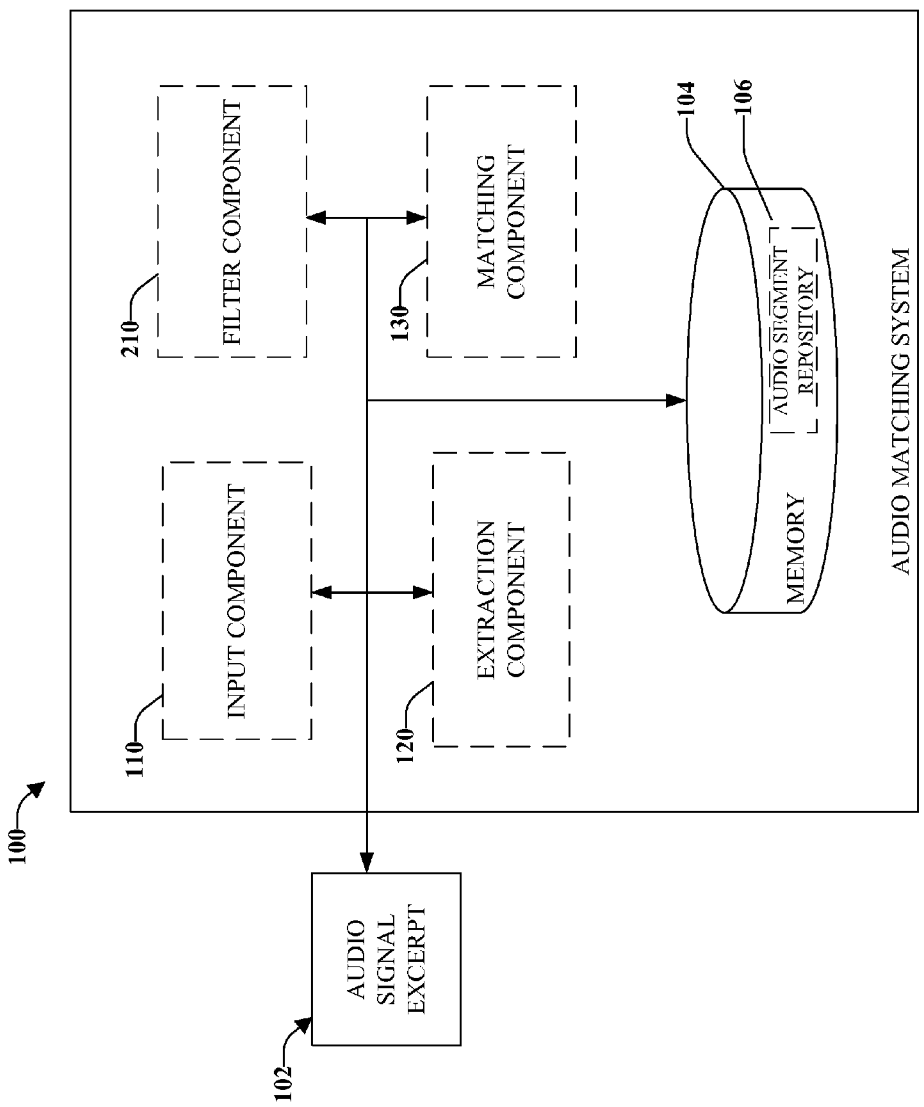

100
AUDIO MATCHING SYSTEM
110
INPUT COMPONENT
120
EXTRACTION COMPONENT
210
FILTER COMPONENT
130
MATCHING COMPONENT
104
106
AUDIO SEGMENT REPOSITORY
MEMORY
102
AUDIO SIGNAL EXCERPT

FIG. 2

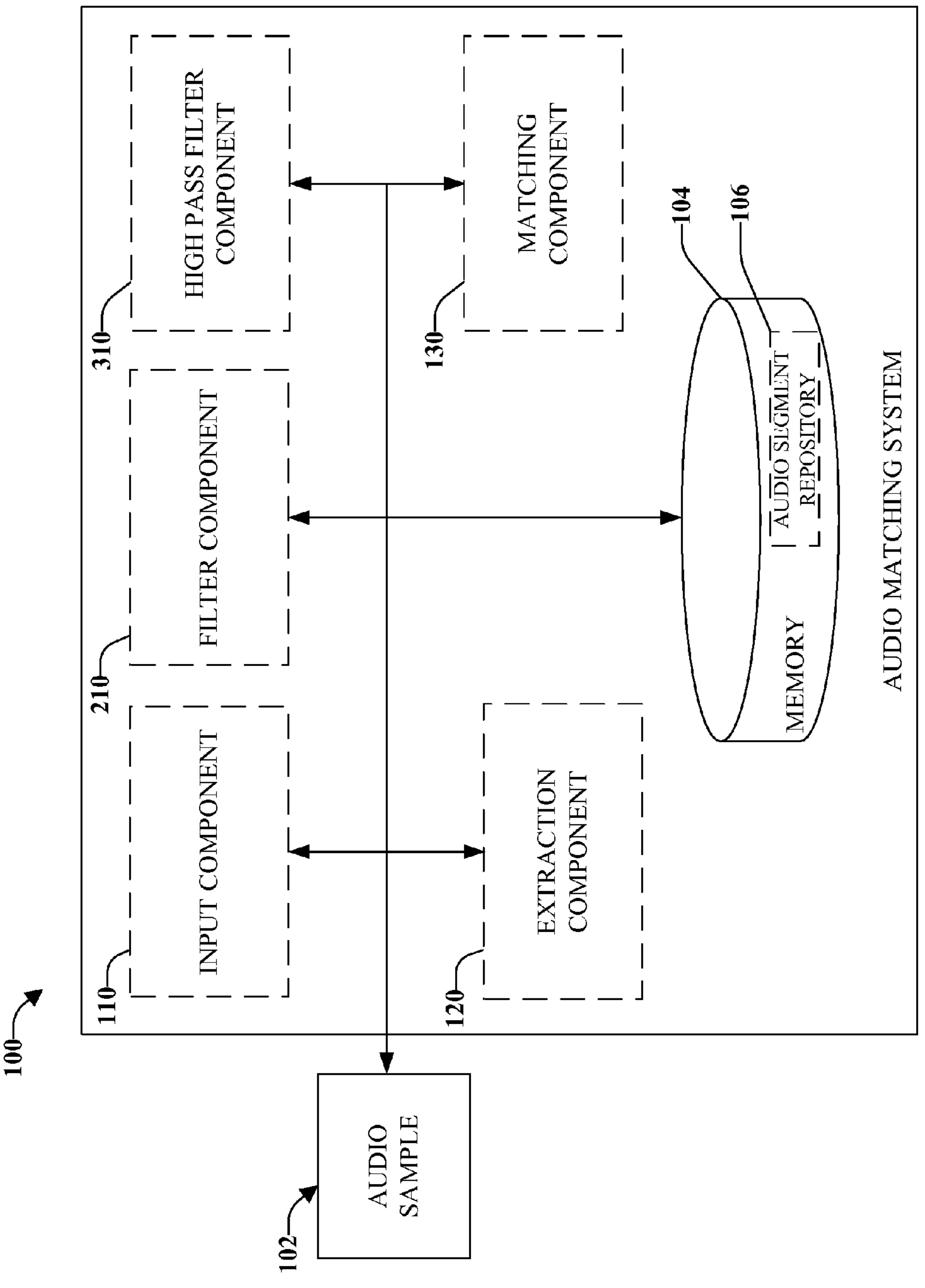

100
110
INPUT COMPONENT
210
FILTER COMPONENT
310
HIGH PASS FILTER COMPONENT
120
EXTRACTION COMPONENT
130
MATCHING COMPONENT
104
106
MEMORY
AUDIO SEGMENT REPOSITORY
AUDIO MATCHING SYSTEM
102
AUDIO SAMPLE

FIG. 3

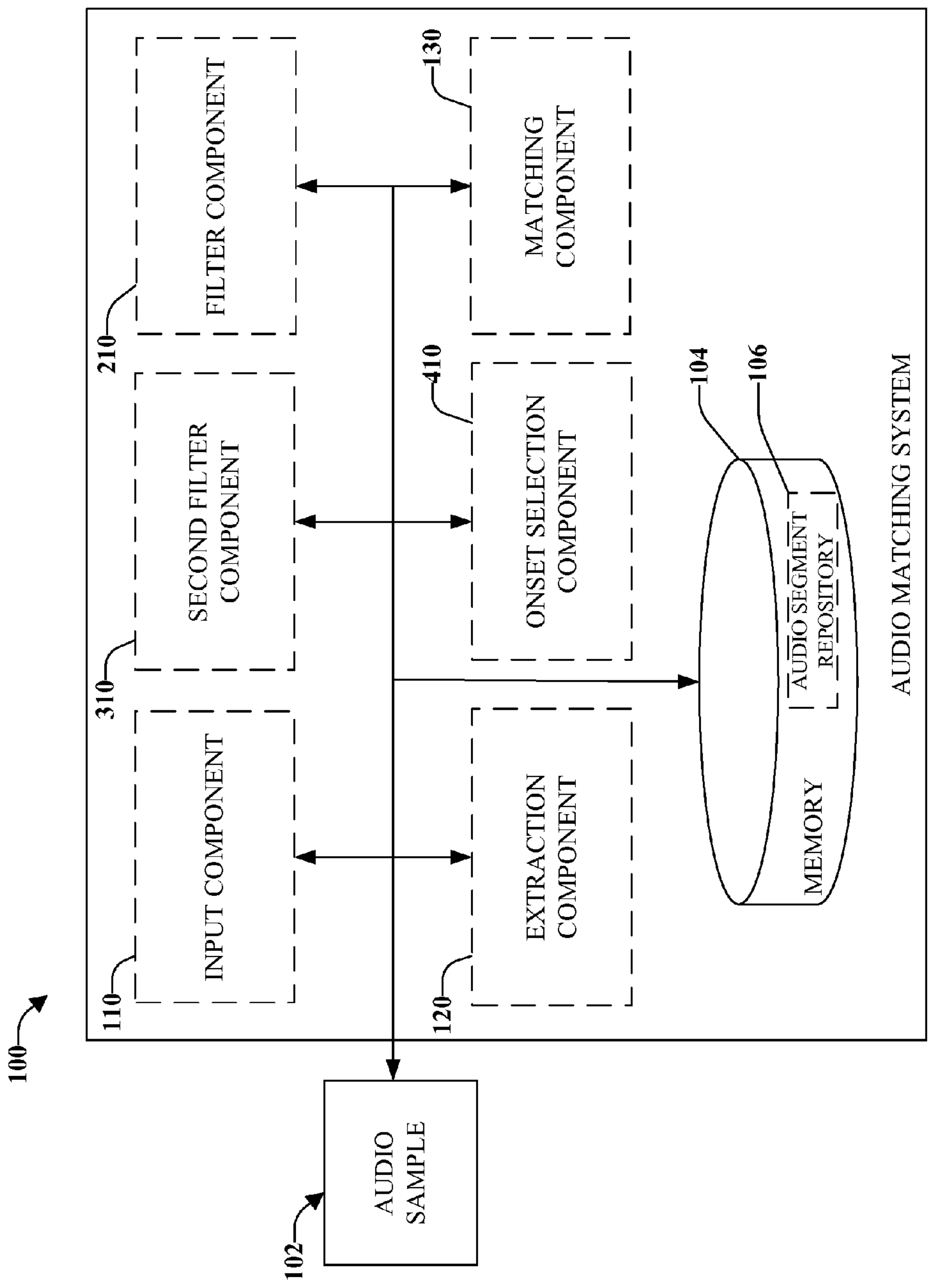

100
AUDIO MATCHING SYSTEM
110
INPUT COMPONENT
310
SECOND FILTER COMPONENT
210
FILTER COMPONENT
120
EXTRACTION COMPONENT
410
ONSET SELECTION COMPONENT
130
MATCHING COMPONENT
104
106
AUDIO SEGMENT REPOSITORY
MEMORY
102
AUDIO SAMPLE

FIG. 4

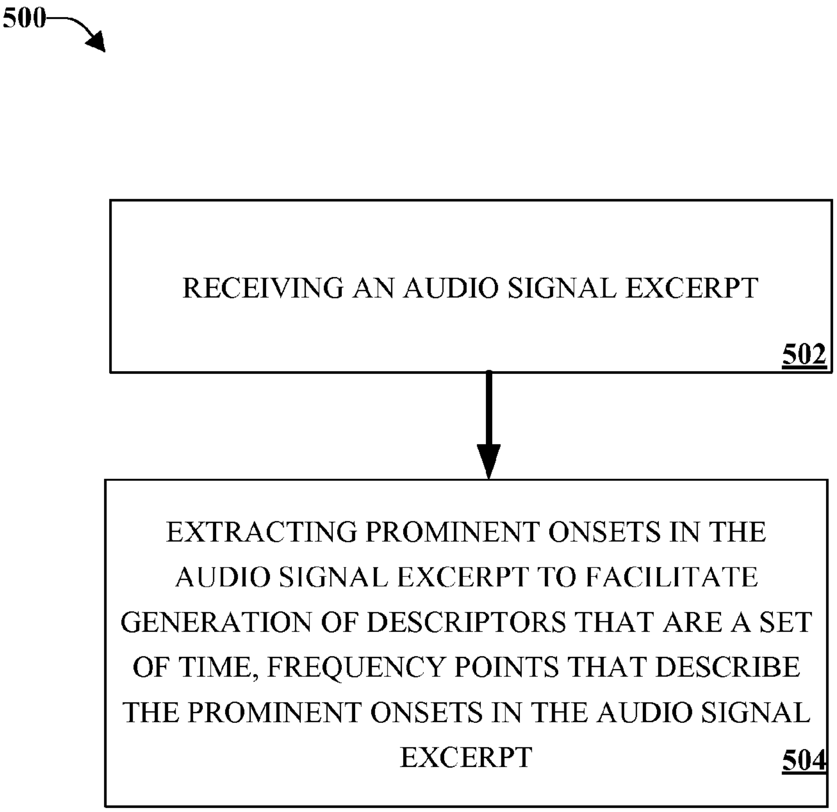

500
RECEIVING AN AUDIO SIGNAL EXCERPT
502
EXTRACTING PROMINENT ONSETS IN THE AUDIO SIGNAL EXCERPT TO FACILITATE GENERATION OF DESCRIPTORS THAT ARE A SET OF TIME, FREQUENCY POINTS THAT DESCRIBE THE PROMINENT ONSETS IN THE AUDIO SIGNAL EXCERPT
504

FIG. 5

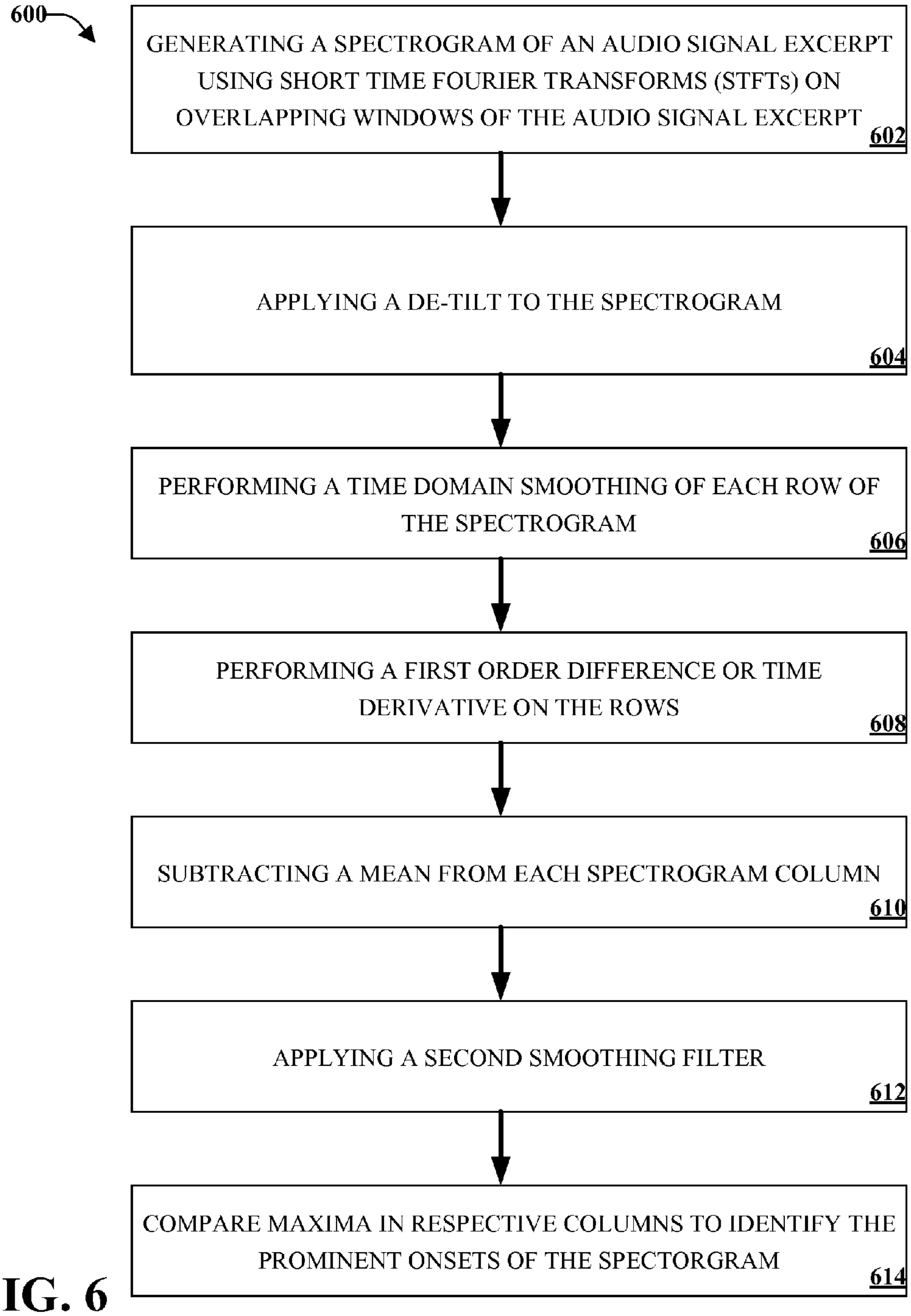

600
GENERATING A SPECTROGRAM OF AN AUDIO SIGNAL EXCERPT USING SHORT TIME FOURIER TRANSFORMS (STFTs) ON OVERLAPPING WINDOWS OF THE AUDIO SIGNAL EXCERPT
602
APPLYING A DE-TILT TO THE SPECTROGRAM
604
PERFORMING A TIME DOMAIN SMOOTHING OF EACH ROW OF THE SPECTROGRAM
606
PERFORMING A FIRST ORDER DIFFERENCE OR TIME DERIVATIVE ON THE ROWS
608
SUBTRACTING A MEAN FROM EACH SPECTROGRAM COLUMN
610
APPLYING A SECOND SMOOTHING FILTER
612
COMPARE MAXIMA IN RESPECTIVE COLUMNS TO IDENTIFY THE PROMINENT ONSETS OF THE SPECTORGRAM
614

FIG. 6

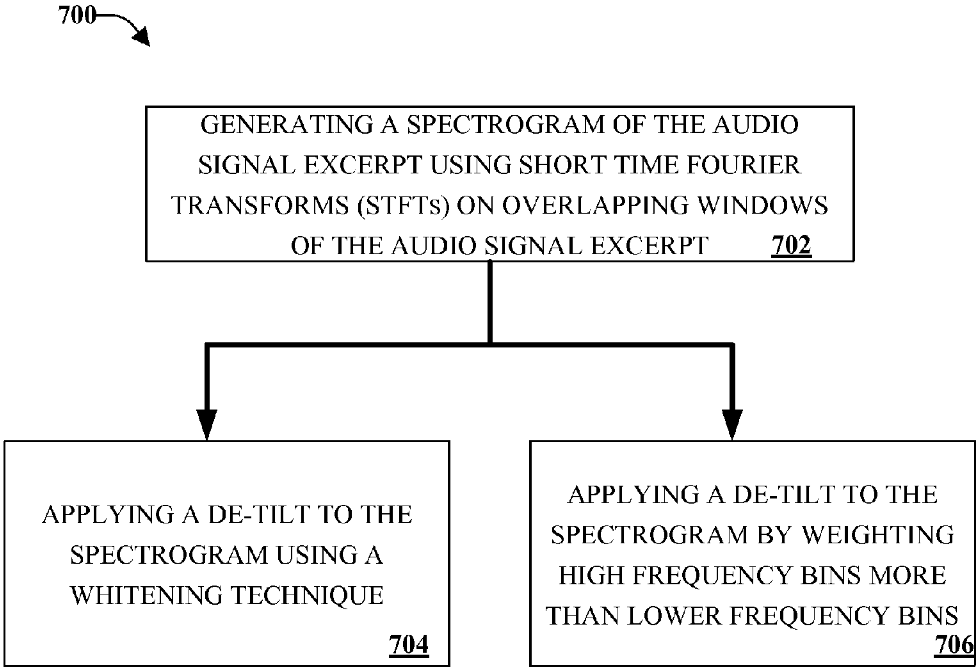

700
GENERATING A SPECTROGRAM OF THE AUDIO SIGNAL EXCERPT USING SHORT TIME FOURIER TRANSFORMS (STFTs) ON OVERLAPPING WINDOWS OF THE AUDIO SIGNAL EXCERPT 702
APPLYING A DE-TILT TO THE SPECTROGRAM USING A WHITENING TECHNIQUE 704
APPLYING A DE-TILT TO THE SPECTROGRAM BY WEIGHTING HIGH FREQUENCY BINS MORE THAN LOWER FREQUENCY BINS 706

FIG. 7

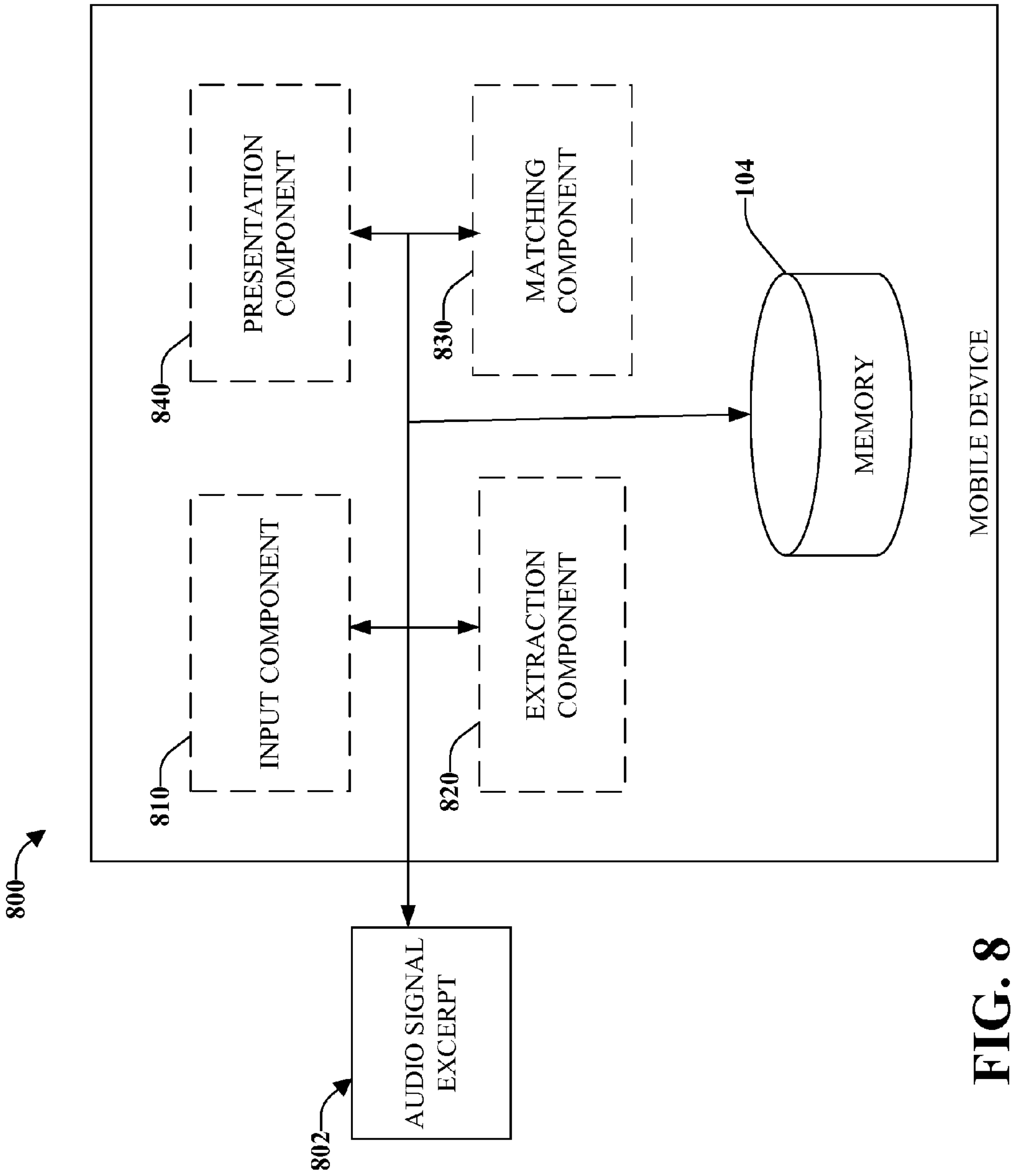

800
PRESENTATION COMPONENT
840
MATCHING COMPONENT
830
104
MEMORY
INPUT COMPONENT
810
EXTRACTION COMPONENT
820
MOBILE DEVICE
AUDIO SIGNAL EXCERPT
802

FIG. 8

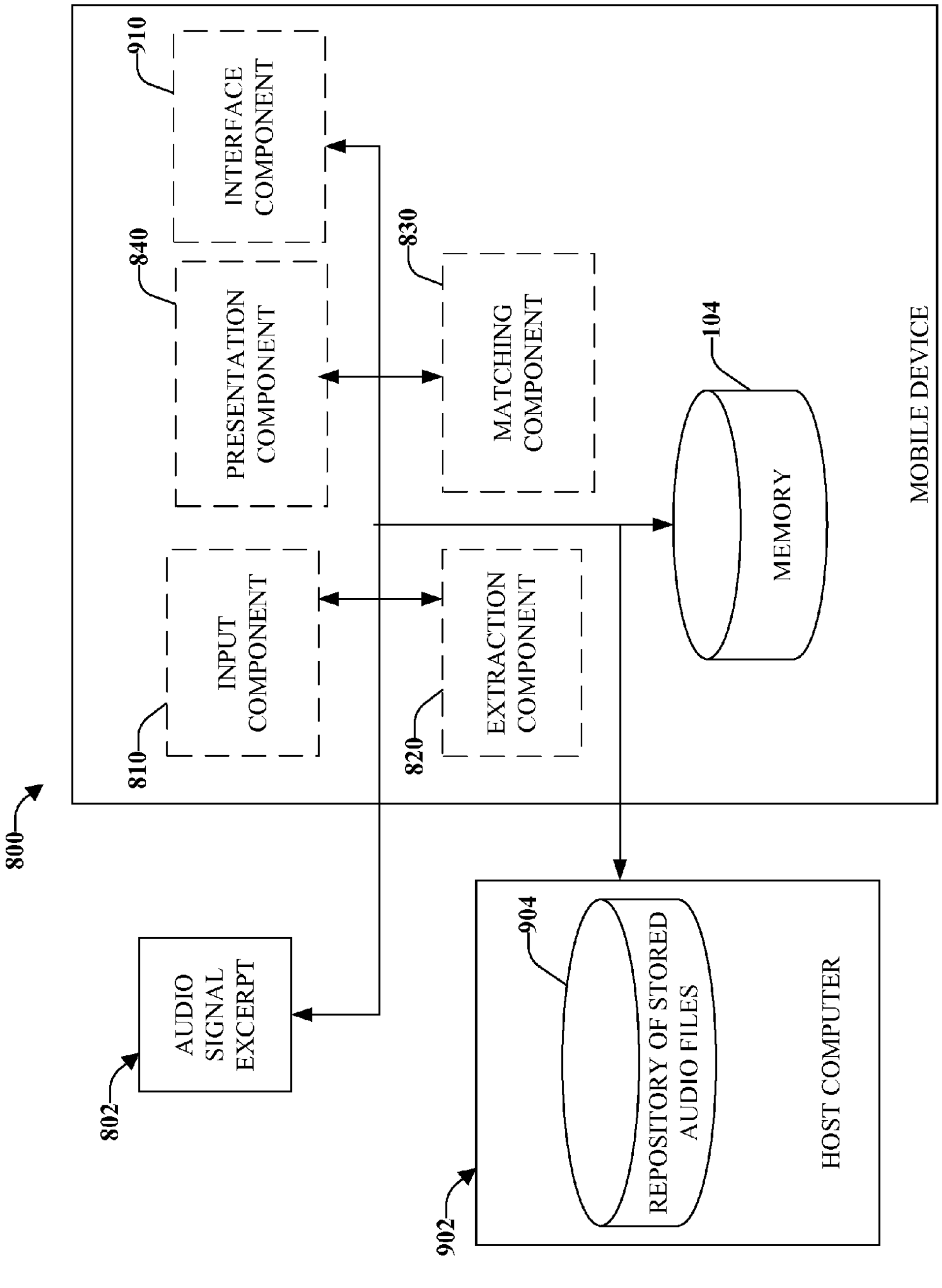

800
MOBILE DEVICE
910
INTERFACE COMPONENT
840
PRESENTATION COMPONENT
830
MATCHING COMPONENT
104
MEMORY
810
INPUT COMPONENT
820
EXTRACTION COMPONENT
802
AUDIO SIGNAL EXCERPT
902
HOST COMPUTER
904
REPOSITORY OF STORED AUDIO FILES

FIG. 9

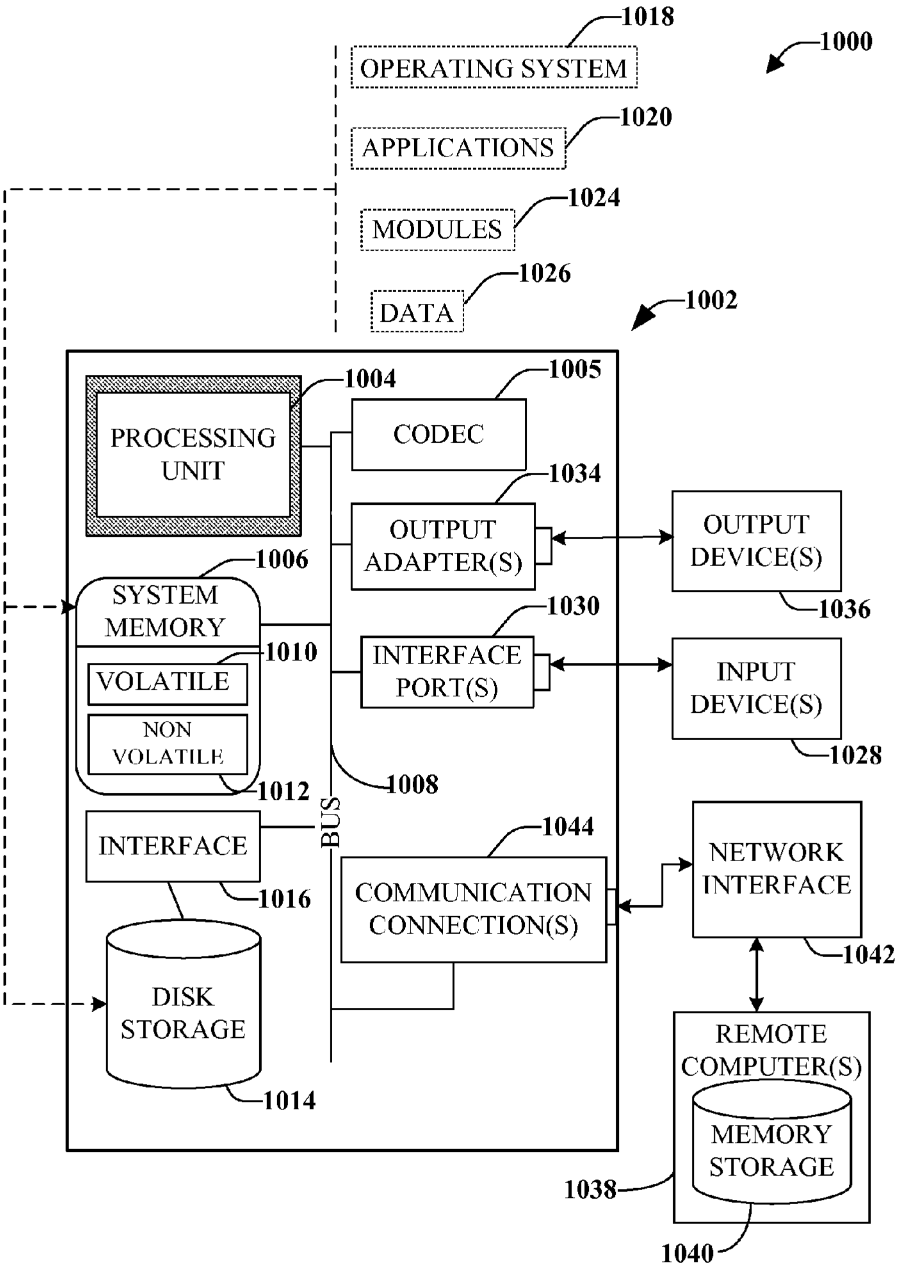

1018
OPERATING SYSTEM
1000
APPLICATIONS
1020
MODULES
1024
DATA
1026
1002
PROCESSING UNIT
1004
CODEC
1005
OUTPUT ADAPTER(S)
1034
OUTPUT DEVICE(S)
1036
SYSTEM MEMORY
1006
VOLATILE
1010
NON VOLATILE
1012
INTERFACE PORT(S)
1030
INPUT DEVICE(S)
1028
1008
BUS
INTERFACE
1016
COMMUNICATION CONNECTION(S)
1044
NETWORK INTERFACE
1042
DISK STORAGE
1014
REMOTE COMPUTER(S)
MEMORY STORAGE
1038
1040

FIG. 10

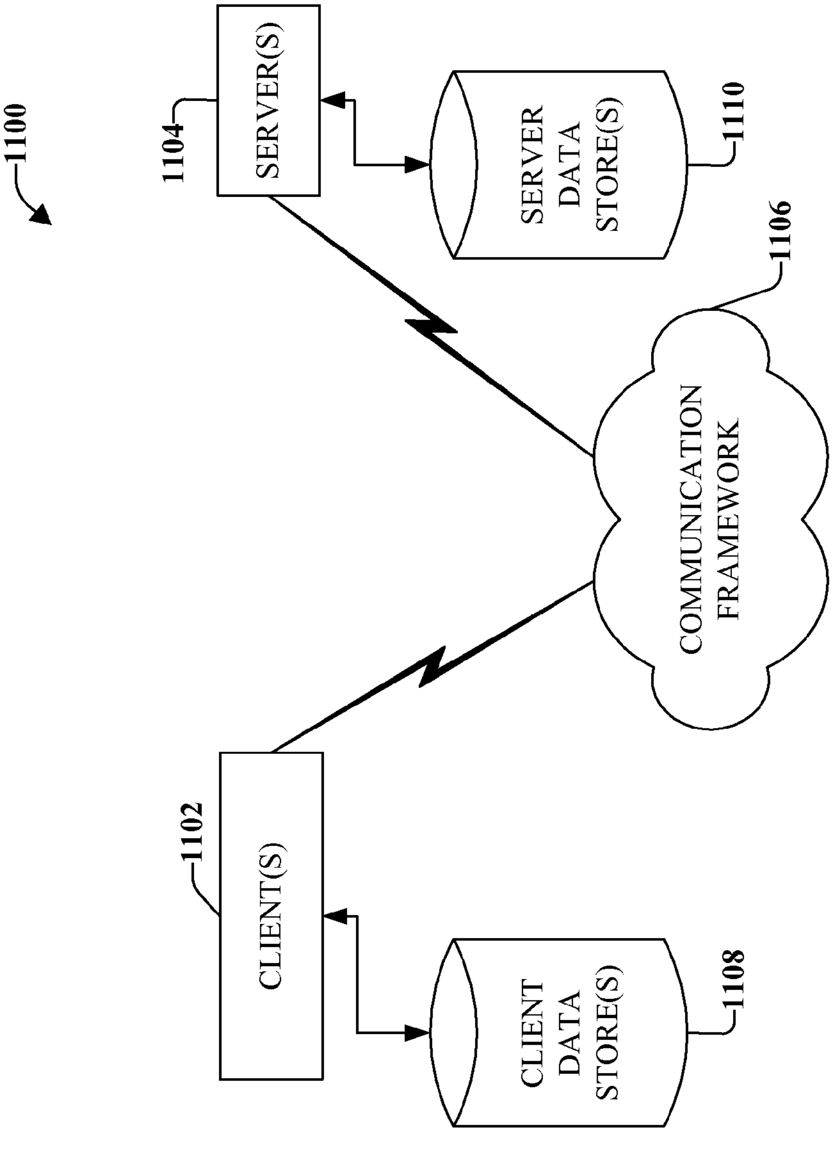

1100
1104
SERVER(S)
SERVER DATA STORE(S)
1110
1106
COMMUNICATION FRAMEWORK
1102
CLIENT(S)
CLIENT DATA STORE(S)
1108

FIG. 11

AUDIO MATCHING USING TIME-FREQUENCY ONSETS

TECHNICAL FIELD

This application relates to audio matching, and more particularly to detecting time-frequency onsets and using the detected onsets to match an audio signal.

BACKGROUND

Audio samples can be recorded by many commercially available electronic devices such as smart phones, tablets, e-readers, computers, personal digital assistants, personal media players, etc. Audio matching provides for identification of a recorded audio sample by comparing the audio sample to a set of reference samples. To make the comparison, an audio sample can be transformed to a time-frequency representation of the sample by using, for example, a short time Fourier transform (STFT). Using the time-frequency representation, interest points that characterize time and frequency locations of peaks or other distinct patterns of the spectrogram can then be extracted from the audio sample. Fingerprints or descriptors can be computed as functions of sets of interest points. Fingerprints of the audio sample can then be compared to fingerprints of reference samples to determine identity of the audio sample.

Recorded audio signals can suffer from many types of distortion. A signal could suffer from noise distortion, pitch shift distortion, compression algorithm distortion, etc. These distortions will affect which interest points are selected. A fingerprint containing distorted interest points may prevent an audio signal from being matched with a fingerprint containing clean undistorted interest points of the same audio signal.

An audio matching system that is robust to distortion is desirable. By generating fingerprints of an audio sample containing interest points that are robust to distortion, the accuracy of an audio matching system relying on the generated fingerprints can be improved. Thus, effectively selecting interest points that are robust to distortion and generating fingerprints based on those interest points is desirable.

SUMMARY

The following presents a simplified summary of the specification in order to provide a basic understanding of some aspects of the specification. This summary is not an extensive overview of the specification. It is intended to neither identify key or critical elements of the specification nor delineate the scope of any particular embodiments of the specification, or any scope of the claims. Its sole purpose is to present some concepts of the specification in a simplified form as a prelude to the more detailed description that is presented in this disclosure.

Systems and methods disclosed herein relate to audio matching. An input component can receive an audio signal excerpt. An extraction component can extract prominent onsets in the audio signal excerpt to facilitate generation of descriptors that are a set of time-frequency points that describe the prominent onsets in the audio signal excerpt.

The following description and the drawings set forth certain illustrative aspects of the specification. These aspects are indicative, however, of but a few of the various ways in which the principles of the specification may be employed.

Other advantages and novel features of the specification will become apparent from the following detailed description of the specification when considered in conjunction with the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. **1** a high-level functional block diagram of an example audio matching system using the detection of time-frequency onsets in accordance with implementations of this disclosure;

FIG. **2** illustrates a high-level functional block diagram of an example audio matching system using the detection of time-frequency onsets and a filter component in accordance with implementations of this disclosure;

FIG. **3** illustrates a high-level functional block diagram of an example audio matching system using the detection of time-frequency onsets and a second filter component in accordance with implementations of this disclosure;

FIG. **4** illustrates a high-level functional block diagram of an example audio matching system using the detection of time-frequency onsets and a high pass filter component in accordance with implementations of this disclosure;

FIG. **5** illustrates an example method for audio matching using the detection of time-frequency onsets in accordance with implementations of this disclosure;

FIG. **6** illustrates an example method for detection of time-frequency onsets in accordance with implementations of this disclosure;

FIG. **7** illustrates an example method for applying a de-tilt to a spectrogram in accordance with implementations of this disclosure;

FIG. **8** illustrates a high-level functional block diagram of an example mobile device capable of matching an audio signal using the detection of time-frequency onsets in accordance with implementations of this disclosure;

FIG. **9** illustrates a high-level functional block diagram of an example mobile device capable of matching an audio signal using the detection of time-frequency onsets including an interface component in accordance with implementations of this disclosure;

FIG. **10** illustrates an example block diagram of a computer operable to execute the disclosed architecture in accordance with implementations of this disclosure; and

FIG. **11** illustrates an example schematic block diagram for a computing environment in accordance with implementations of this disclosure.

DETAILED DESCRIPTION

The innovation is now described with reference to the drawings, wherein like reference numerals are used to refer to like elements throughout. In the following description, for purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of this innovation. It may be evident, however, that the innovation can be practiced without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to facilitate describing the innovation.

Audio matching in general involves analyzing an audio sample for unique characteristics that can be used in comparison to unique characteristics of reference samples to identify the audio sample. As a starting point for this analysis, a spectrogram can be constructed of the audio signal. A spectrogram represents an audio sample by plotting time on one axis and frequency on another axis. Additionally, amplitude or intensity of a certain frequency at a certain time can also be incorporated into the spectrogram by using color or a third dimension.

There are several different techniques for creating a spectrogram. One technique involves using a series of band-pass filters that can filter an audio sample at a specific frequency and measure amplitude of the audio sample at that specific frequency over time. The audio sample can be run through additional filters to individually isolate a set of frequencies to measure the amplitude of the set over time. A spectrogram can be created by combining all the measurements over time on the frequency axis to generate a spectrogram image of frequency amplitudes over time.

A second technique involves using short-time Fourier transform ("STFT") to break down an audio sample into time windows, where each window is Fourier transformed to calculate a magnitude of the frequency spectrum for the duration of each window. Combining a plurality of windows side by side on the time axis of the spectrogram creates an image of frequency amplitudes over time. It can be appreciated that time windows can be overlapping. Other techniques, such as wavelet transforms, can also be used to construct a spectrogram.

Creating and storing in a database an entire spectrogram for a plurality of reference samples can require large amounts of storage space and affect scalability of an audio matching system. Therefore, it is desirable to instead calculate and store compact descriptors ("fingerprints") of reference samples versus an entire spectrogram. One method of calculating fingerprints is to first determine individual interest points that identify unique characteristics of local features of the time-frequency representation of the reference sample. Fingerprints can then be computed as functions of sets of interest points.

Calculating interest points involves identifying unique characteristics of the spectrogram. For example, an interest point could be a spectral peak of a specific frequency over a specific window of time. As another non-limiting example, an interest point could also include timing of the onset of a note. It is to be appreciated that conceivably any suitable spectral event over a specific duration of time could constitute an interest point.

Generating interest points corresponding to a time-frequency onset of a spectral event can provide benefits in that onsets can be highly robust to transformations such as noise distortion, pitch shift distortion, or compression distortion. An audio matching system based on generating descriptors containing time-frequency onsets can provide similar advantages in regards to more accurately identifying audio signals suffering from distortions.

Systems and methods herein provide for generating interest points that correspond to time-frequency onsets. A spectrogram can be generated based on an audio signal excerpt using STFTs on overlapping windows. A de-tilt can be applied by, for example, using a whitening technique or weighting STFT windows so that high frequency bins are weighted more than low frequency bins. A time-domain smoothing can be applied to each row of the spectrogram where rows correspond to frequency bands. A first-order difference or time derivative can be applied to each row where the difference in consecutive windows (e.g., a low value followed by a high value) can signify an onset. A mean can be subtracted from each column of the spectrogram where columns correspond to time bands. A second smoothing filter can then be applied. The maximum value for each time column can be compared to neighboring bins wherein if that maximum value is higher than the maxima in the two neighboring columns, it can signify an onset. Descriptors can then be generated as functions of sets of time-frequency onsets. It can be appreciated that other information related to characterizing the onset can be included in a descriptor, such as, for example, a duration or time span of the note relating to the onset, neighboring spectrogram energy levels, or other local features relating to the onset.

As discussed in greater detail below, various implementations provide for using descriptors containing time-frequency onsets as interest points in an audio matching system. An input component can receive an audio signal excerpt. An extraction component can extract prominent onsets in the audio signal excerpt to facilitate generation of descriptors that are sets of time-frequency data that describe the prominent onsets in the audio signal excerpt.

Referring to FIG. **1**, there is illustrated a high-level functional block diagram of an example audio matching system using the detection of time-frequency onsets in accordance with implementations of this disclosure. Input component **110** can receive an audio signal excerpt **102**. Extraction component **120** can extract prominent onsets in the audio signal excerpt **102** to facilitate generation of descriptors that are a set of time-frequency data that describe the prominent onsets in audio signal excerpt **102**.

In one implementation, extraction component **120** can compute descriptors as a function of sets of interest points that are local features of a time-frequency data, the time-frequency data being a representation of the audio signal excerpt, such as, for example, a spectrogram. The function of sets of interest points can be STFTs.

In one implementation, extraction component **120** can extract prominent onsets that are robust to noise distortion, pitch shift distortion, compression distortions, or other distortions.

In one implementation, extraction component **120** can generate a spectrogram of audio signal excerpt **102** using STFTs on overlapping windows of the audio signal. For example, if a window of 10 milliseconds is chosen as a STFT window, then each window (e.g., a column when time is represented on a horizontal axis of the spectrogram) will overlap both a preceding window and a following window. It can be appreciated that in alternate implementations, differing window sizes can be employed.

In one implementation, extraction component **120** can apply a de-tilt to the spectrogram. Generally, there is more energy in lower frequency bands. Adjusting for this additional energy can aid in identifying time-frequency onsets. The de-tilt can be applied using a whitening technique such as, for example, using a whitening transform, or it can be applied by weighting STFT windows or bins such that high frequency bins are weighted more than lower frequency bins.

A matching component **130** that matches the audio signal excerpt to an audio signal stored in a repository to generate a set of potential audio matches wherein the matching component further employs the descriptors to identify a match of the audio signal excerpt to one of the potential audio matches among the set of potential audio matches as a function of overlap of prominent onsets. It can be appreciated that repository **106** can be disparately located from system **100**. In one implementation, audio segment repository **106** can contain descriptors of audio segments. The descriptors can then be used by matching component **130** to match the audio signal excerpt **102**. In one implementation, after receiving a set of potential matching reference signals based on the descriptors contained within audio segment repository **106**, matching component **130** can match the audio signal excerpt **102** and the set of potential matching reference signals as a function of overlap of prominent onsets. The more overlap of prominent onsets, the more likely the audio signal excerpt **102** is a match for one of the potential matching reference signals in the set of potential matching reference signals.

Referring to FIG. **2**, there is illustrated a high-level functional block diagram of an example audio matching system using the detection of time-frequency onsets including a filter component **210** that can emphasize onsets by performing a time-domain smoothing of each row of the spectrogram. In one implementation, rows of the spectrogram correspond to frequency bands. In one implementation, filter component **210** can emphasize fluctuations within a range of time scales associated with the prominent onsets.

In one implementation, filter component **210** can be a two-pole critically damped filter. A smoothing time constant can be selected, for example, 50 ms. It can be appreciated that in other implementations a different time constant can be selected and used in conjunction with a two-pole critically damped filter.

Referring to FIG. **3**, there is illustrated a high-level functional block diagram of an example audio matching system using the detection of time-frequency onsets and a high pass filter component **310** that can perform a first-order difference on each row of the spectrogram or a time derivative on each row of the spectrogram. It can be appreciated that a large difference in consecutive time windows or bins, e.g., a low value followed by a high value, can be indicative of a time-frequency onset. In one implementation, high pass filter component **310** can subtract a mean from each spectrogram column. The mean can be the mean related to spectral activity over a certain number of windows or a specific time frame including over the duration of the entire audio signal excerpt. High pass filter component **310** can also apply a second time domain smoothing to the spectrogram. In another implementation, high pass filter component **310** can work in conjunction with filter component **210** to apply a second time domain smoothing to the spectrogram.

Referring to FIG. **4**, there is illustrated a high-level functional block diagram of an example audio matching system using the detection of time-frequency onsets including an onset selection component **410** that can identify for each column or time step in the spectrogram a highest valued bin, and compare maxima in neighboring columns to identify prominent time-frequency onsets of the spectrogram. For example, if the maximum value of a time step (column) is higher than both neighboring time steps then it is likely a prominent time-frequency onset.

FIGS. **5-7** illustrate methods and/or flow diagrams in accordance with this disclosure. For simplicity of explanation, the methods are depicted and described as a series of acts. However, acts in accordance with this disclosure can occur in various orders and/or concurrently, and with other acts not presented and described herein. Furthermore, not all illustrated acts may be required to implement the methods in accordance with the disclosed subject matter. In addition, those skilled in the art will understand and appreciate that the methods could alternatively be represented as a series of interrelated states via a state diagram or events. Additionally, it should be appreciated that the methods disclosed in this specification are capable of being stored on an article of manufacture to facilitate transporting and transferring such methods to computing devices. The term article of manufacture, as used herein, is intended to encompass a computer program accessible from any computer-readable device or storage media.

Moreover, various acts have been described in detail above in connection with respective system diagrams. It is to be appreciated that the detailed description of such acts in the prior figures can be and are intended to be implementable in accordance with the following methods.

FIG. **5** illustrates an example method for audio matching using the detection of time-frequency onsets in accordance with implementations of this disclosure. At **502**, an audio signal excerpt can be received (e.g., by an input component). At **504**, prominent onsets in the audio signal excerpt can be extracted (e.g., by an extraction component) to facilitate generation of descriptors that are a set of time-frequency data that describe the prominent onsets in the audio signal excerpt.

FIG. **6** illustrates an example method for detection of time-frequency onsets in accordance with implementations of this disclosure. At **602**, a spectrogram can be generated (e.g. by an input component) of an audio signal excerpt using STFTs on overlapping windows of the audio signal excerpt. At **604**, a de-tilt can be applied (e.g., by an extraction component). At **606**, a time domain smoothing can be performed (e.g., by a filter component). At **608** a first order difference or time derivative can be performed (e.g., by a high pass filter component) on each row of the spectrogram. It can be appreciated that a row can be a frequency band. At **610** a mean can be subtracted (e.g., by a high-pass filter component). At **612**, a second smoothing filter can be applied (e.g., by a filter component). At **614**, maxima in respective columns can be compared (e.g., by an onset selection component) to identify prominent onsets of the spectrogram.

FIG. **7** illustrates an example method for applying a de-tilt to a spectrogram in accordance with implementations of this disclosure. At **702**, a spectrogram can be generated (e.g. by an input component) of an audio signal excerpt using STFTs on overlapping windows of the audio signal excerpt. At **704**, a de-tilt can be applied (e.g., by an extraction component) to the spectrogram using a whitening technique. Alternatively, at **706**, a de-tilt can be applied (e.g., by an extraction component) to the spectrogram by weighting high frequency bins more than lower frequency bins.

Referring to FIG. **8**, there is illustrated a high-level functional block diagram of an example mobile device capable of matching an audio signal using the detection of time-frequency onsets in accordance with implementations of this disclosure. An example of a mobile device can include smart phones, tablets, e-readers, computers, personal digital assistants, personal media players, etc. The mobile device can contain a microphone capable of receiving an audio signal. An input component **810** can receive an audio signal excerpt of the audio signal.

An extraction component **820** can extract prominent onsets in the audio signal excerpt to facilitate generation of descriptors that are a set of time-frequency data that describe the prominent onsets in the audio signal excerpt. A matching component **830** can employ the descriptors to identify a match of the audio signal excerpt. A presentation component **840** can notify a user of the mobile device identification of the matched audio file stored in the repository. For example, presentation component can display the identification on the mobile device, audibly identify the identification, send email identification to the user, send an SMS message regarding the identification to the user, etc.

Referring to FIG. **9**, there is illustrated a high-level functional block diagram of an example mobile device capable of matching an audio signal using the detection of time-frequency onsets including an interface component **910** that communicatively couples the matching component **830** to a repository of stored audio files **904**, wherein the repository of stored audio files **904** is located within a host computer **902**.

In one implementation, the matching component can perform the match by transmitting the descriptors to the host computer **802**, wherein the host computer **802** matches the uploaded descriptors to a set of stored descriptors stored in the repository as a function of overlap of prominent onsets. It can be appreciated that the entirety of each reference audio sample need not be stored within the repository.

Reference throughout this specification to "one implementation," or "an implementation," means that a particular feature, structure, or characteristic described in connection with the implementation is included in at least one implementation. Thus, the appearances of the phrase "in one implementation," or "in an implementation," in various places throughout this specification can, but are not necessarily, referring to the same implementation, depending on the circumstances. Furthermore, the particular features, structures, or characteristics may be combined in any suitable manner in one or more implementations.

To the extent that the terms "includes," "including," "has," "contains," variants thereof, and other similar words are used in either the detailed description or the claims, these terms are intended to be inclusive in a manner similar to the term "comprising" as an open transition word without precluding any additional or other elements.

As used in this application, the terms "component," "module," "system," or the like are generally intended to refer to a computer-related entity, either hardware (e.g., a circuit), software, a combination of hardware and software, or an entity related to an operational machine with one or more specific functionalities. For example, a component may be, but is not limited to being, a process running on a processor (e.g., digital signal processor), a processor, an object, an executable, a thread of execution, a program, and/or a computer. By way of illustration, both an application running on a controller and the controller can be a component. One or more components may reside within a process and/or thread of execution and a component may be localized on one computer and/or distributed between two or more computers. A filter component can be hardware capable of filtering. Further, a "device" can come in the form of specially designed hardware; generalized hardware made specialized by the execution of software thereon that enables hardware to perform specific functions (e.g. generating interest points and/or fingerprints); software on a computer readable medium; or a combination thereof.

The aforementioned systems, circuits, modules, and so on have been described with respect to interaction between several components and/or blocks. It can be appreciated that such systems, circuits, components, blocks, and so forth can include those components or specified sub-components, some of the specified components or sub-components, and/or additional components, and according to various permutations and combinations of the foregoing. Sub-components can also be implemented as components communicatively coupled to other components rather than included within parent components (hierarchical). Additionally, it should be noted that one or more components may be combined into a single component providing aggregate functionality or divided into several separate sub-components, and any one or more middle layers, such as a management layer, may be provided to communicatively couple to such sub-components in order to provide integrated functionality. Any components described herein may also interact with one or more other components not specifically described herein but known by those of skill in the art.

Moreover, the words "example" or "exemplary" are used herein to mean serving as an example, instance, or illustration. Any aspect or design described herein as "exemplary" is not necessarily to be construed as preferred or advantageous over other aspects or designs. Rather, use of the words "example" or "exemplary" is intended to present concepts in a concrete fashion. As used in this application, the term "or" is intended to mean an inclusive "or" rather than an exclusive "or". That is, unless specified otherwise, or clear from context, "X employs A or B" is intended to mean any of the natural inclusive permutations. That is, if X employs A; X employs B; or X employs both A and B, then "X employs A or B" is satisfied under any of the foregoing instances. In addition, the articles "a" and "an" as used in this application and the appended claims should generally be construed to mean "one or more" unless specified otherwise or clear from context to be directed to a singular form.

With reference to FIG. **10**, a suitable environment **1000** for implementing various aspects of the claimed subject matter includes a computer **1002**. The computer **1002** includes a processing unit **1004**, a system memory **1006**, a codec **1005**, and a system bus **1008**. The system bus **1008** couples system components including, but not limited to, the system memory **1006** to the processing unit **1004**. The processing unit **1004** can be any of various available processors. Dual microprocessors and other multiprocessor architectures also can be employed as the processing unit **1004**.

The system bus **1008** can be any of several types of bus structure(s) including the memory bus or memory controller, a peripheral bus or external bus, and/or a local bus using any variety of available bus architectures including, but not limited to, Industrial Standard Architecture (ISA), Micro-Channel Architecture (MSA), Extended ISA (EISA), Intelligent Drive Electronics (IDE), VESA Local Bus (VLB), Peripheral Component Interconnect (PCI), Card Bus, Universal Serial Bus (USB), Advanced Graphics Port (AGP), Personal Computer Memory Card International Association bus (PCMCIA), Firewire (IEEE 1394), and Small Computer Systems Interface (SCSI).

The system memory **1006** includes volatile memory **1010** and non-volatile memory **1012**. The basic input/output system (BIOS), containing the basic routines to transfer information between elements within the computer **1002**, such as during start-up, is stored in non-volatile memory **1012**. By way of illustration, and not limitation, non-volatile memory **1012** can include read only memory (ROM), programmable ROM (PROM), electrically programmable ROM (EPROM), electrically erasable programmable ROM (EEPROM), or flash memory. Volatile memory **1010** includes random access memory (RAM), which acts as external cache memory. According to present aspects, the volatile memory may store the write operation retry logic (not shown in FIG. **10**) and the like. By way of illustration and not limitation, RAM is available in many forms such as static RAM (SRAM), dynamic RAM (DRAM), synchronous DRAM (SDRAM), double data rate SDRAM (DDR SDRAM), enhanced SDRAM (ESDRAM).

Computer **1002** may also include removable/non-removable, volatile/non-volatile computer storage media. FIG. **10** illustrates, for example, a disk storage **1014**. Disk storage **1014** includes, but is not limited to, devices like a magnetic disk drive, solid state disk (SSD) floppy disk drive, tape drive, Jaz drive, Zip drive, LS-100 drive, flash memory card, or memory stick. In addition, disk storage **1014** can include storage media separately or in combination with other storage media including, but not limited to, an optical disk drive such as a compact disk ROM device (CD-ROM), CD recordable drive (CD-R Drive), CD rewritable drive (CD-RW Drive) or a digital versatile disk ROM drive (DVD-ROM). To facilitate connection of the disk storage devices **1014** to the system bus **1008**, a removable or non-removable interface is typically used, such as interface **1016**.

It is to be appreciated that FIG. **10** describes software that acts as an intermediary between users and the basic computer resources described in the suitable operating environment **1000**. Such software includes an operating system **1018**. Operating system **1018**, which can be stored on disk storage **1014**, acts to control and allocate resources of the computer system **1002**. Applications **1020** take advantage of the management of resources by operating system **1018** through program modules **1024**, and program data **1026**, such as the boot/shutdown transaction table and the like, stored either in system memory **1006** or on disk storage **1014**. It is to be appreciated that the claimed subject matter can be implemented with various operating systems or combinations of operating systems.

A user enters commands or information into the computer **1002** through input device(s) **1028**. Input devices **1028** include, but are not limited to, a pointing device such as a mouse, trackball, stylus, touch pad, keyboard, microphone, joystick, game pad, satellite dish, scanner, TV tuner card, digital camera, digital video camera, web camera, and the like. These and other input devices connect to the processing unit **1004** through the system bus **1008** via interface port(s) **1030**. Interface port(s) **1030** include, for example, a serial port, a parallel port, a game port, and a universal serial bus (USB). Output device(s) **1036** use some of the same type of ports as input device(s) **1028**. Thus, for example, a USB port may be used to provide input to computer **1002**, and to output information from computer **1002** to an output device **1036**. Output adapter **1034** is provided to illustrate that there are some output devices **1036** like monitors, speakers, and printers, among other output devices **1036**, which require special adapters. The output adapters **1034** include, by way of illustration and not limitation, video and sound cards that provide a means of connection between the output device **1036** and the system bus **1008**. It should be noted that other devices and/or systems of devices provide both input and output capabilities such as remote computer(s) **1038**.

Computer **1002** can operate in a networked environment using logical connections to one or more remote computers, such as remote computer(s) **1038**. The remote computer(s) **1038** can be a personal computer, a server, a router, a network PC, a workstation, a microprocessor based appliance, a peer device, a smart phone, a tablet, or other network node, and typically includes many of the elements described relative to computer **1002**. For purposes of brevity, only a memory storage device **1040** is illustrated with remote computer(s) **1038**. Remote computer(s) **1038** is logically connected to computer **1002** through a network interface **1042** and then connected via communication connection(s) **1044**. Network interface **1042** encompasses wire and/or wireless communication networks such as local-area networks (LAN) and wide-area networks (WAN) and cellular networks. LAN technologies include Fiber Distributed Data Interface (FDDI), Copper Distributed Data Interface (CDDI), Ethernet, Token Ring and the like. WAN technologies include, but are not limited to, point-to-point links, circuit switching networks like Integrated Services Digital Networks (ISDN) and variations thereon, packet switching networks, and Digital Subscriber Lines (DSL).

Communication connection(s) **1044** refers to the hardware/software employed to connect the network interface **1042** to the bus **1008**. While communication connection **1044** is shown for illustrative clarity inside computer **1002**, it can also be external to computer **1002**. The hardware/software necessary for connection to the network interface **1042** includes, for exemplary purposes only, internal and external technologies such as, modems including regular telephone grade modems, cable modems and DSL modems, ISDN adapters, and wired and wireless Ethernet cards, hubs, and routers.

Referring now to FIG. **11**, there is illustrated a schematic block diagram of a computing environment **1100** in accordance with the subject specification. The system **1100** includes one or more client(s) **1102**, which can include an application or a system that accesses a service on the server **1104**. The client(s) **1102** can be hardware and/or software (e.g., threads, processes, computing devices). The client(s) **1102** can house cookie(s), metadata, and/or associated contextual information by employing the specification, for example.

The system **1100** also includes one or more server(s) **1104**. The server(s) **1104** can also be hardware or hardware in combination with software (e.g., threads, processes, computing devices). The servers **1104** can house threads to perform, for example, extracting, applying a de-tilt, performing time domain smoothing, performing time derivates, etc. in accordance with the subject disclosure. One possible communication between a client **1102** and a server **1104** can be in the form of a data packet adapted to be transmitted between two or more computer processes where the data packet contains, for example, an audio sample or descriptors associated with an audio sample. The data packet can include a cookie and/or associated contextual information, for example. The system **1100** includes a communication framework **1106** (e.g., a global communication network such as the Internet) that can be employed to facilitate communications between the client(s) **1102** and the server(s) **1104**.

Communications can be facilitated via a wired (including optical fiber) and/or wireless technology. The client(s) **1102** are operatively connected to one or more client data store(s) **1108** that can be employed to store information local to the client(s) **1102** (e.g., cookie(s) and/or associated contextual information). Similarly, the server(s) **1104** are operatively connected to one or more server data store(s) **1110** that can be employed to store information local to the servers **1104**.

The illustrated aspects of the disclosure may also be practiced in distributed computing environments where certain tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules can be located in both local and remote memory storage devices.

The systems and processes described above can be embodied within hardware, such as a single integrated circuit (IC) chip, multiple ICs, an application specific integrated circuit (ASIC), or the like. Further, the order in which some or all of the process blocks appear in each process should not be deemed limiting. Rather, it should be understood that some of the process blocks can be executed in a variety of orders that are not all of which may be explicitly illustrated herein.

What has been described above includes examples of the implementations of the present invention. It is, of course, not possible to describe every conceivable combination of components or methods for purposes of describing the claimed subject matter, but many further combinations and permutations of the subject innovation are possible. Accordingly, the claimed subject matter is intended to embrace all such alterations, modifications, and variations that fall within the spirit and scope of the appended claims. Moreover, the above description of illustrated implementations of this disclosure, including what is described in the Abstract, is not intended to be exhaustive or to limit the disclosed implementations to the precise forms disclosed. While specific implementations and examples are described herein for illustrative purposes, various modifications are possible that are considered within the scope of such implementations and examples, as those skilled in the relevant art can recognize.

In particular and in regard to the various functions performed by the above described components, devices, circuits, systems and the like, the terms used to describe such components are intended to correspond, unless otherwise indicated, to any component which performs the specified function of the described component (e.g., a functional equivalent), even though not structurally equivalent to the disclosed structure, which performs the function in the herein illustrated exemplary aspects of the claimed subject matter. In this regard, it will also be recognized that the innovation includes a system as well as a computer-readable storage medium having computer-executable instructions for performing the acts and/or events of the various methods of the claimed subject matter.

What is claimed is:

1. A system comprising:

a memory that has stored thereon computer executable components; and

at least one processor that executes the following computer executable components stored in the memory:

an input component that receives an audio signal excerpt; and

an extraction component that:

generates a spectrogram of the audio signal excerpt using short time Fourier transforms (STFTs), wherein rows in the spectrogram represent frequencies and columns in the spectrogram represent times;

applies a first smoothing filter to each row of the spectrogram to produce a smoothed spectrogram, wherein the first smoothing filter is a time-domain smoothing filter;

performs a time derivative on each row of the smoothed spectrogram to produce a derived spectrogram;

subtracts a mean from each spectrogram column of the derived spectrogram to produce a subtracted spectrogram;

applies a second smoothing filter to the subtracted spectrogram to produce a second smoothed spectrogram, wherein the second smoothing filter is different than the first smoothing filter;

extracts prominent onsets in the audio signal excerpt based on the second smoothed spectrogram to facilitate generation of descriptors that are based on a set of time-frequency data that describe the prominent onsets in the audio signal; and

a matching component that matches the audio signal excerpt to an audio signal stored in a repository to generate a set of potential audio matches wherein the matching component further employs the descriptors to identify a match of the audio signal excerpt to one of the potential audio matches among the set of potential audio matches as a function of overlap of prominent onsets.

2. The system of claim **1**, wherein the descriptors are a function of sets of interest points that are local features of a time-frequency data, the time-frequency data being a representation of the audio signal excerpt.

3. The system of claim **1**, wherein the extraction component applies a de-tilt to the spectrogram to produce a de-tilted spectrogram, and wherein the first smoothing filter is applied to the de-tilted spectrogram.

4. The system of claim **3**, wherein the de-tilt is performed by weighting STFT bins so that high frequency bins are weighted more than lower frequency bins.

5. The system of claim **3**, wherein the de-tilt is performed using a whitening technique.

6. The system of claim **1**, comprising a filter component that emphasizes fluctuations within a range of time scales associated with the prominent onsets.

7. The system of claim **6**, wherein the filter component is a two-pole critically damped filter.

8. The system of claim **1**, further comprising an onset selection component that identifies for each column in the second smoothed spectrogram a highest value frequency bin, and compares the respective highest value frequency bins in immediately neighboring columns to identify the prominent onsets of the spectrogram.

9. A method, comprising:

receiving, by a system including a processor, an audio signal excerpt;

generating, by the system, a spectrogram of the audio signal excerpt using short time Fourier transforms (STFTs), wherein rows in the spectrogram represent frequencies and columns in the spectrogram represent times;

applying, by the system, a first smoothing filter to each row of the spectrogram to produce a smoothed spectrogram, wherein the first smoothing filter is a time-domain smoothing filter;

performing, by the system, a time derivative on each row of the spectrogram to produce a derived spectrogram;

subtracting, by the system, a mean from each spectrogram column of the derived spectrogram to produce a subtracted spectrogram;

applying, by the system, a second smoothing filter to the subtracted spectrogram to produce a second smoothed spectrogram, wherein the second smoothing filter is different than the first smoothing filter;

extracting, by the system, prominent onsets in the audio signal excerpt based on the second smoothed spectrogram to facilitate generation of descriptors that are a set of time-frequency data that describe the prominent onsets in the audio signal excerpt;

comparing the audio signal excerpt to a plurality of audio signals stored in a repository to generate a set of potential matches; and

matching the audio signal excerpt to a potential match among the set of potential matches as a function of overlap of prominent onsets.

10. The method of claim **9**, wherein the generating the spectrogram of the audio signal excerpt using the short time Fourier transforms (STFTs) comprises employing overlapping windows of the audio signal excerpt.

11. The method of claim **10**, comprising applying a de-tilt to the spectrogram to produce a de-tilted spectrogram, and wherein the first smoothing filter is applied to the de-tilted spectrogram.

12. The method of claim **11**, wherein applying the de-tilt uses a whitening technique.

13. The method of claim **11**, wherein applying the de-tilt weights respective short time Fourier transform bins wherein weights of high frequency bins are greater than weights of low frequency bins.

14. The method of claim **10**, wherein the descriptors are a function of sets of interest points that are local features of a time-frequency data, the time-frequency data being a representation of the audio signal excerpt.

15. The method of claim **9**, comprising emphasizing fluctuations within a range of time scales associated with the prominent onsets.

16. The method of claim **14**, further comprising:

identifying for each column in the second smoothed spectrogram a highest value frequency bin; and

comparing the respective highest value frequency bins in immediately neighboring columns; and

determining that each highest value frequency bin in a column that has a higher value than respective highest value frequency bins in the immediately neighboring columns is a potential prominent onset.

17. A system, comprising:

means for receiving an audio signal excerpt;

means for generating a spectrogram of the audio signal excerpt using short time Fourier transforms (STFTs), wherein rows in the spectrogram represent frequencies and columns in the spectrogram represent times;

means for applying a first smoothing filter to each row of the spectrogram to produce a smoothed spectrogram, wherein the first smoothing filter is a time-domain smoothing filter;

means for performing time derivative on each row of the smoothed spectrogram to produce a derived spectrogram;

means for subtracting a mean from each spectrogram column of the derived spectrogram to produce a subtracted spectrogram;

means for applying a second smoothing filter to the subtracted spectrogram to produce a second smoothed spectrogram, wherein the second smoothing filter is different than the first smoothing filter; and

means for extracting prominent onsets in the audio signal excerpt based on the second smoothed spectrogram to facilitate generation of descriptors that are a set of time-frequency data that describe the prominent onsets in the audio signal excerpt; and

means for employing the descriptors to identify a match of the audio signal excerpt to an audio file stored in a repository as a function of overlap of prominent onsets.

18. A mobile device, comprising:

a microphone for receiving an audio signal;

a memory that has stored thereon computer executable components; and

a processor that executes the following computer executable components stored in the memory:

an input component that receives an audio signal excerpt of the audio signal;

an extraction component that:

generates a spectrogram of the audio signal excerpt using short time Fourier transforms (STFTs), wherein rows in the spectrogram represent frequencies and columns in the spectrogram represent times;

applies a first smoothing filter to each row of the spectrogram to produce a smoothed spectrogram, wherein the first smoothing filter is a time-domain smoothing filter;

performs a time derivative on each row of the smoothed spectrogram to produce a derived spectrogram;

subtracts a mean from each spectrogram column of the derived spectrogram to produce a subtracted spectrogram;

applies a second smoothing filter to the subtracted spectrogram to produce a second smoothed spectrogram, wherein the second smoothing filter is different than the first smoothing filter; and

extracts prominent onsets in the audio signal excerpt based on the second smoothed spectrogram to facilitate generation of descriptors that are a set of time-frequency data that describe the prominent onsets in the audio signal excerpt;

a matching component that employs the descriptors to identify a match of the audio signal excerpt and an audio file stored in a repository; and

a presentation component that notifies a user of the mobile device identification of the matched audio file stored in the repository.

19. The mobile device of claim **18**, further comprising an onset selection component that identifies for each column in the second smoothed spectrogram a highest value frequency bin, and compares the respective highest value frequency bins in immediately neighboring columns to identify the prominent onsets of the spectrogram.

20. The mobile device of claim **19**, wherein the onset selection component compares the respective highest value frequency bins in immediately neighboring columns including determining that a highest value frequency bin in a column that is higher than respective highest value frequency bin in the immediately neighboring columns is a prominent onset.

21. The system of claim **8**, wherein the onset selection component compares the respective highest value frequency bins in immediately neighboring columns including determining that a highest value frequency bin in a column that has a higher value than respective highest value frequency bins in the immediately neighboring columns is a prominent onset.

22. The method of claim **9**, wherein the comparing the respective highest value frequency bins in immediately neighboring columns includes determining that a highest value frequency bin in a column that has a higher value than respective highest value frequency bins in the immediately neighboring columns is a prominent onset.

23. The system of claim **17**, further comprising:

means for identifying for each column in the second smoothed spectrogram a highest value frequency bin; and

means for comparing the respective highest value frequency bins in immediately neighboring columns; and

means for determining that each highest value frequency bin in a column that has a higher value than respective highest value frequency bins in the immediately neighboring columns is a potential prominent onset.

* * * * *